Automatic discovery of Latin syntactic changes

Micha Elsner and **Emily Lane**
melsner0@gmail and lane.434@osu.edu
Department of Linguistics
The Ohio State University

Abstract

Syntactic change tends to affect constructions, but treebanks annotate lower-level structure: PCFG rules or dependency arcs. This paper extends prior work in native language identification, using Tree Substitution Grammars to discover constructions which can be tested for historical variability. In a case study comparing Classical and Medieval Latin, the system discovers several constructions corresponding to known historical differences, and learns to distinguish the two varieties with high accuracy. Applied to an intermediate text (the Vulgate Bible), it indicates which changes between the eras were already occurring at this earlier stage.

1 Introduction

In recent years, the study of language variation and change has been aided by a variety of computational tools that can automatically infer hypotheses about language change from a corpus (Eisenstein, 2015). In the domain of syntax, however, computational work is still limited by the necessity of manually choosing interesting hypotheses to study. For example, computational research on the syntax of African-American English (Stewart, 2014) is driven by pre-existing scholarly intuitions about the distinctive features of this dialect, but such intuitions are much harder to obtain for dead (or newly-emerging) language varieties.

This paper adopts a method for unsupervised learning of syntactic constructions previously found effective for native language identification (Swanson and Charniak, 2012), and shows that it can discover a range of historically varying elements in a Latin corpus. In particular, we conduct a case study comparing classical prose (1st century CE) with the Medieval writing of Thomas Aquinas (c. 1270) and the intermediate stage of the Vulgate Bible (4th century CE). Such a method can be used for the initial "hypothesis discovery" step in a historical research project. Although the method is currently incapable of discovering some (lexically bound) constructions, we demonstrate that it discovers several interpretable and interesting historical changes.

The method (which we review more fully below) induces a Tree Substitution Grammar (TSG) from a constituency treebank. TSG rules are larger than Context-Free Grammar (CFG) rules and thus have the power to represent constructions, including partial lexicalization. We use chi-squared feature selection to rank the TSG rules for their sensitivity to historical change. We evaluate the rules both by building classifiers to identify the historical period of unknown text, and by manual examination and interpretation.

2 Variationist research

Computational methods for studying language variation can enhance both diachronic (historical) and synchronic (sociolinguistic) research. In some cases, the computational contribution is to build a classifier for a particular feature which is already of interest. For instance, Bane et al. (2010) target pre-selected phonetic features for analysis in recorded speech. Other computational systems are exploratory: capable of discovering new hypotheses about geographical or social variation in the data. But existing systems of this type are lexicographic. For instance, Eisenstein (2015) detects previously unknown local slang terms, such as "deadass" in New York City. Rao et al. (2010) discover words and ngrams correlated with gender and other social attributes, as do later papers such as Bamman et al. (2014).

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 156–164,
Berlin, Germany, August 11, 2016.

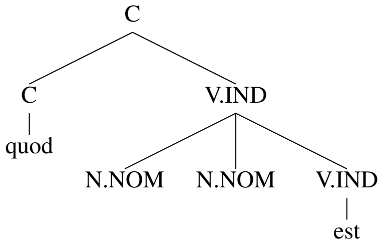

Figure 1: A TSG fragment with the root symbol *C* (complement clause), introducing an indicative subclause headed by *quod* which contains two nominals and the verb *est* "is".

Work on syntactic variation is much rarer. For the most part, it is confirmatory rather than explanatory; computational systems are designed to find examples of specific constructions in order to support investigations driven by pre-existing hypotheses. Such systems do not suggest new hypotheses from the data. Stewart (2014) detects African-American copula deletion and auxiliary verb structures; Doyle (2014) investigates "needs done" and double modals. We know of one exploratory project using syntactic features: Johannsen et al. (2015) use universal dependencies to extract "treelets" correlated with age and gender. Our TSG fragments are similar to their treelet features, but have the potential to be larger and are partly lexicalized.

3 Tree substitution grammars

A Tree Substitution Grammar (TSG) generalizes Context-Free Grammar (CFG) by allowing rules to insert arbitrarily large tree *fragments* (Cohn et al., 2009). Each fragment has a root symbol (analogous to the left-hand-side category in a CFG) and a *frontier* which can consist of terminals (words) and non-terminal symbols to be filled in later in the derivation. An example tree fragment is shown in figure 1; this fragment describes a particular complement clause structure which can be interpreted as the construction "that X is Y".

A single treebank tree may have multiple TSG derivations (depending on how it is split up into constructional fragments), so TSGs must be induced from the data. The Data-oriented Parsing (DOP) method (Bod and Kaplan, 1998) was criticized by Johnson (2002) for its poor estimation procedure. Newer methods select a set of fragments either using Bayesian models (Cohn et al., 2009; Post and Gildea, 2009) or using so-called Double-DOP (Sangati and Zuidema, 2011), which creates a TSG rule for every maximal fragment which occurs more than once in the dataset. For instance, the fragment in figure 1 would be extracted from the trees for *dicit quod Cicero consul est* and *quod Caesar dux est scimus*,[1] since it is shared between them both, but cannot be further expanded without adding an unshared element. TSGs are equivalent in expressive power to CFGs and can be efficiently parsed using the same algorithms (Goodman, 1996).

TSGs have been used effectively for native language identification (Swanson and Charniak, 2012): determining the native language of a writer with intermediate proficiency in English, given a sample of their English writing. (Two closely related approaches are Wong and Dras (2011) and Wong et al. (2012).) Swanson and Charniak (2014) show that the rules learned by their system can be interpreted as transferring features or constructions from their native language. In this work, we argue that TSG is also useful for detecting the forms of change which occur in historical corpora.

4 Classical and Medieval Latin

Lind (1941) divides Latin roughly into Classical (250 BCE to 100 CE), Late (100-600 CE), Medieval (600-1300) and Neo-Latin (1300-1700). Though these divisions are heuristic, they do correspond to episodes of lexical and grammatical change. Medieval Latin was an educated language used by clerics and scholars. It diverges from its Classical roots partly due to the influence of the evolving Romance languages and of Church texts (themselves often influenced by Hebrew and Greek) (Lind, 1941; Löfstedt, 1959).

Scholars debate the nature and origins of variability within Medieval Latin. Löfstedt (1959, ch. 3) surveys this research. For instance, an early theory that African Late Latin was syntactically distinct was rejected on the grounds that the supposedly African constructions represented a distinct rhetorical style rather than a dialect. Similar questions have been raised about dialectal differences between France and Spain and the influences of Germanic languages on their local varieties of Latin.

A robust computational method could help to resolve controversies like these. In many cases, the dispute is centered around some construction

[1] "He says Cicero is consul" and "That Caesar is a general, we know".

which is claimed to be a regional variant. For instance, Hanssen (1945) claims that *mittere pro* may be a calque of English "send for", a claim which Löfstedt (1959) rebuts by providing a variety of examples from elsewhere. The constructions involved may be quite rare, and a specialist in one region or period may be unaware that a construction of interest is attested elsewhere, especially in obscure texts. An automatic method for discovering cases which vary across regions or periods could not only help to reject this type of spurious claim, but also find genuine examples of regional variation which may not have been previously noticed.

5 Case study

To demonstrate the effectiveness of our method, we use it to construct a classifier which differentiates between single utterances of Classical and Medieval Latin. The classifier features are a set of TSG fragments. We induce the TSG from training data, then run a feature selection procedure to limit their number. We show that the learned classifier is fairly effective, and analyze two sets of its learned features by hand, connecting them to the literature on known historical changes.

As a secondary question, we investigate the placement of the Vulgate Bible: is it more similar to Classical or Medieval Latin? The Vulgate is often seen as an intermediate between the two periods. Sidwell (1995, p.30) says that it:

> "sanctified usages such as changes in the use of cases and the subjunctive, and the more frequent use of *quod/quia* clauses in reported speech. . . . It is linguistically a central text."

But while the Vulgate has a strong influence on Medieval tradition, its compiler, St. Jerome,[2] was classically educated; in a famous letter, he actually chastised himself for being "a Ciceronian, not a Christian" (Wright, 1933) because of his preference for classical prose over the "uncultivated" Biblical style. Running the classifier on sentences from the Vulgate can reveal how the text balances these two affinities.

Author	Text	Sents.	Date
Classical (Perseus)			
Cicero	In Catalinam	327	63 BCE
Sallust	Bellum Catalinae	701	c. 42 BCE
Caesar	de Bello Gallico	71	c. 57 BCE
Petronius	Satyricon	1114	c. 54-68 CE
Late (Perseus)			
Jerome (editor)	Vulgate Bible (Revelation)	405	c. 380 CE
Medieval (Thomisticus)			
Thomas Aquinas	Summa Contra Gentiles	9859	c. 1250-70

Table 1: Authors and texts used in the current study; dates from (Shipley et al., 2008; Hornblower et al., 2012).

6 Data and preprocessing

Our case study uses two Latin treebanks, Perseus (Bamman and Crane, 2011) and Index Thomisticus (Passarotti, 2011), each of which contains dependency-parsed Latin prose (Bamman et al., 2007). Table 1 provides a list of authors, dates and sizes. Unfortunately, the Late and Medieval groups are represented by a single author each; this represents a weakness of this project, since it will be impossible to distinguish Medieval Latin in general from the specific style of Aquinas. The data is also somewhat unbalanced, with Aquinas representing much more text than any other author. These limitations are imposed by the system's requirement for parse trees, and the unavailability of other parsed Latin data.

Both source treebanks use non-projective dependency trees. To employ the TSG technique, we convert these to constituency trees. Our conversion introduces a phrasal projection over every head word with children; following Klein and Manning (2004), we give this projection the same label as the head word's part of speech. Non-projective edges are converted to projective ones by reordering the words so that the descendants of every head are contiguous. When a subtree is moved for this purpose, its tag is marked with a diacritic, so that the grammar can learn separate

[2] Our sample, the book of Revelation, was "slightly revised" (Sidwell, 1995) by Jerome from an older Latin translation of the 2nd century CE (Hornblower et al., 2012).

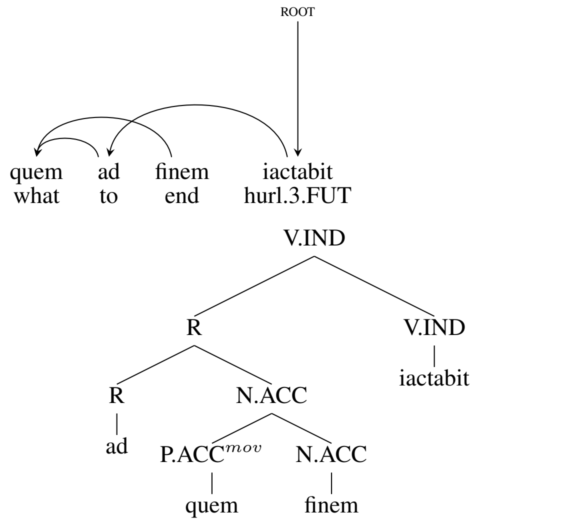

Figure 2: Transformation of the non-projective construction *quem ad finem iactabit* ("how far will [your audacity] hurl [itself]") from the Perseus Treebank into a constituency structure. ^{mov} is the "moved element" diacritic.

rules for non-projective constructions.[3] See figure 2 for an example.

The treebanks use multidimensional part of speech tags, with slightly different tagging conventions. We use only the top-level part of speech tag for most words, converting the Thomisticus tags deterministically into the Perseus tags. We use the remaining dimensions to annotate nominals with their case and verbs with their mood (indicative, subjunctive, imperative or infinitive).

Finally, again following Swanson and Charniak (2013), we selectively delexicalize the trees. This prevents the "syntactic" patterns our system learns as markers of variation from being dominated by lexical items marking different topics (Sarawgi et al., 2011). For instance, Aquinas frequently uses the adjective *Christiana* "Christian", while the Classical authors do not. But this is a change in culture, rather than in language.

We remove all lexical items except prepositions (POS tag *R*), conjunctions (*C*), and a short list of adverbials (*D*), *ne, non, tam, tamen, ita, etiam* ("lest, not, so, however, thus, besides"). We replace all forms of the verb *esse* ("to be"), which is often used as an auxiliary, with the Perseus Treebank lemmatized form *sum1*. For example, the phrase *ad quem finem* is delexicalized to *ad UNK UNK*, although the part of speech tags remain as a guide to the grammatical form.

Finally, we split our data into random train/dev/test sections, with $\frac{1}{10}$ of each era for development, $\frac{2}{10}$ testing and the rest training. Since we do not train or develop on the Vulgate, we set this data aside as a single set.

7 Learning and ranking constructions

We extract a set of TSG rules using Double-DOP (Sangati and Zuidema, 2011). As stated above, this process yields the set of all maximal TSG fragments which occur in more than one treebank tree. It is usually more exhaustive than the Bayesian extractors (Cohn et al., 2009), although it can be slow for large corpora.

As in Swanson and Charniak (2013), we then match each rule against each treebank sentence, deciding whether that rule can occur in any derivation of the sentence. We assemble these decisions into a $(rule \times sentence)$ binary co-occurrence matrix. To compute variants which change between Classical and Medieval Latin, we sum across sentences in the training set to compute the four-way contingency table of counts: sentences with and without the rule in each era. We compute the χ^2 statistic for each table and use this to rank the rules for feature selection. Swanson and Charniak (2013) recommend χ^2 because it tends to retain moderately rare rules with good predictive power, rather than focusing on generally-applicable rules with weak predictions (as is the case for Information Gain).

We select rules for which the χ^2 probability is less than .00001 (tuned on development experiments; the corresponding χ^2 statistic is about 19). In our dataset, this method selects 357 TSG fragments. We use the Megam (Daumé III, 2004) maximum entropy classifier to learn a predictor for the era (Classical or Medieval) of a sentence given the binary feature vector indicating presence or absence of these 357 fragments. Results are shown in table 2. The classifier overpredicts the majority class (Medieval) but still achieves 77% accuracy on the minority class, indicating that its features are reasonably informative about language change.

7.1 Analysis

We will discuss two interpretable patterns discovered by our system: a known historical change in

[3] Unlike the construction of Nivre and Nilsson (2005), this tag is intended to *describe* non-projective constructions but does not give enough information to *parse* them.

Era	N	Correct	Acc
Classical	442	341	77%
Medieval	1972	1931	98%
Majority	2414	1972	82%
Overall	2414	2268	94%

Table 2: Classifier accuracy on test data.

the use of complement clauses, and a stable but hard-to-interpret pattern in adjective/noun ordering. Finally, we discuss our failure to detect the decline of a parenthetical construction called the ablative absolute. In each case, we have manually grouped together TSG fragments selected by the system and imposed an interpretation on them by doing additional linguistic analysis.

Classical Latin verbs like *dicere* "to say" typically take nonfinite complement clauses (Pinkster, 1990). In Medieval Latin, these verbs more commonly take finite complement clauses, often with the complementizer *quod* "that" (Sidwell, 1995, p.368). The two sentences below (the first from Cicero, the second from Aquinas) exemplify these different structures:

(1) Lepidum te habitare
Lepidus.ACC you.ACC live.with.INF
velle dixisti
want.INF say.2.PFV
"You said you want to live with Lepidus."

(2) Dicitur quod sapientia
say.3.PASS COMP wisdom.NOM
infinitus thesaurus est
infinite treasury be.3
"It is said that wisdom is an infinite treasury."

Table 3 shows a collection of tree fragments related to this change, along with their χ^2 statistic values. The system clearly identifies the Medieval complementizer *quod* "that" with both indicative and subjunctive clauses, along with *autem* "however" and *igitur* "therefore". Although these do occur in Classical prose, the high values of the χ^2 statistic show that they are clearly much more widely used in Medieval Latin. The system also identifies the decline of the Classical complementizer *cum* ("when" with indicatives, "since" with subjunctives). The low χ^2 value (46) for the infinitival complement clause, however, must be accounted as a partial failure of the system; this fragment appears low in the selected list of features.

Fragment	χ^2	hits
More Classical		
(V.INF N.ACC V.INF)	46	69
(C (C cum) V.SBJV)	299	68
(C (C cum) V.IND	102	24
More Medieval		
(C igitur)	353	1575
(C (C autem) V.IND)	351	1475
(C (C quod) V.IND)	161	990
(C (C quod) V.SBJV)	150	738

Table 3: Hand-selected features related to changes in complement clauses.

The system's failure to extract this construction with high confidence stems from an inability to generalize over the contents of the subclause. Due to the flat tree structure, a subclause with a temporal modifier, for example: *dico te* ***priore nocte*** *venisse* "I say that you came **last night**" cannot be unified with a subclause without. This leads to data fragmentation, and therefore to a low frequency for the construction, which reduces the system's confidence in associating it with the Classical period.

A second interpretable set of TSG fragments governs adjective ordering. It consists of all the rules *(N.case N.case Adj.case)* (a nominal, headed by a noun with a postnominal adjective) and *(N.case Adj.case N.case)* (prenominal adjective). Table 4 shows the statistics. With the exception of the ablative case, the Classical data slightly prefers postnominal adjectives, while the Medieval data strongly prefers prenominals. Ledgeway (2012) states that Classical Latin used postnominal adjectives in unmarked contexts, with prenominals serving some semantic and pragmatic functions. This preference is claimed to be stable throughout the Middle Ages, leading to modern Romance languages with mainly postnominal adjectives. Our Medieval corpus data does not follow this pattern, since prenominals are more typical. But whether this reflects an actual localized or temporary change, or Aquinas's personal style, cannot be determined without further investigation.

The ablative absolute is an adverbial modifier that is frequently used to denote a time, or the cause of an action, and often takes the place of a subordinate clause. However, the ablative absolute is not grammatically dependent on any word in its sentence (Allen and Greenough, 1983, p. 263).

Case	% postnominals		χ^2
	Classical	Medieval	
Nom	53	26	135
Gen	56	25	115
Dat	65	8	90
Acc	57	34	413
Abl	35	36	228

Table 4: Percent of postnominal adjectives in noun-adjective phrases, and χ^2 value for the postnominal rule. The Classical data contains more postnominals, while the Medieval data contains more prenominals.

It normally consists of a noun and a passive participle, (although another noun or an adjective can replace the participle):

(3) Omni pacata Gallia
All.ABL pacified.PAST.PART Gaul.ABL
ad eos exercitus noster
against them army.NOM our.NOM
adduceretur
lead.3.SBJV
"**With all of Gaul having been pacified**, our army would be led against them"

There is current speculation that the ablative absolute descends from either an instrumental or a locative origin (Allen and Greenough, 1983). Ramat (1991) argues that it developed from a very colloquial style of speech, as a way to compensate for a lack of "complementizers, auxiliaries, and determiners" (p. 261). Furthermore, Ramat argues that the construction is "more pragmatic than syntactic", and thus declined as Medieval Latin became more formal and syntactically rigid.

The system finds several rules for ablative noun/participle phrases, but none with a χ^2 value above 40. We detect 56 uses in the Classics and 65 in the Medieval corpus. This construction is hand-annotated in the treebanks, however, so we can check our accuracy. In fact, the Classical corpus contains 105 ablative absolutes, while the Medieval corpus has none. Our system underdetects the Classical cases due to modifiers and reorderings, as discussed above. It overdetects the Medieval ones; Medieval constructions that appear to be ablative absolutes often contain gerunds rather than passive participles, an issue hidden by delexicalization and the use of coarse tags. Additionally, Thomas favors a construction similar to the ablative absolute, but which is actually a prepositional phrase:

(4) Quem in rebus cognoscendis
That in things.ABL known.PART
quotidie experimur
daily experience.1.PL
"That we experience daily **in the knowing of things**"

Thus, we miss this historical change because the ablative absolute is quite varied in form, and because our representation fails to distinguish it from similar constructions.

7.2 Late Latin: The Vulgate

We run the Classical/Medieval classifier on the Vulgate, with results shown in table 5. Despite the classifier's overall bias towards the Medieval class, we find that the Vulgate is generally more Classical. However, the proportion of sentences labeled in this way (64%) is not comparable to the 77% of Classical sentences labeled as Classical, indicating that the Vulgate is indeed intermediate between the two eras.

To determine which features most typify the Classical and Medieval components of the Vulgate, we compute the summed contribution of each feature to the entire set of decisions. If a feature f_i has weight θ_i, we compute its importance $M(i)$ over a set of examples x:

$$M(i) = \sum_x |f_i \theta_i| \qquad (1)$$

The top 5 features for each class are shown in table 6. Several features represent changes in adjective ordering (discussed above) and the use of complementizers or clause-initial markers. A few, such as the occurrence of conjunctions and adverbs, do not represent real historical changes and are presumably markers of specific topics or styles. The importance attached to the preposition *in* may reflect either a stylistic difference, or the Medieval tendency to use a preposition where Classical Latin uses the bare ablative case (Sidwell, 1995, p. 367). We believe these results show that the system can aid a linguist in finding language change, but that the output still needs to be analyzed and interpreted by hand.

With the exception of *cum*, the clausal features discussed above have little impact on the classification of Vulgate sentences. To determine whether

	N	%
Total	405	100
Labeled Classical	258	64
Labeled Medieval	147	36

Table 5: Classifier results on the Late Latin Vulgate.

Classical	$M(i)$	Medieval	$M(i)$
Postnominal adj. (abl)	868	Genitive pronouns	757
Any conjunction	768	Preposition *in* "in"	713
Preposition *super* "on"	725	Clause-initial *et* "and"	601
Postnominal adj. (acc)	631	Any adverb	559
Conjunction *cum* "when"	600	Postnominal adjective in PP	507

Table 6: Features important in the classification of Vulgate sentences, ranked by importance $M(i)$.

this represents a failure to generalize, or genuine ambiguity, we search the Vulgate Book of Revelations by hand for verbs with clausal complements; these are not particularly frequent, accounting for their small importance weights. However, both types of complements appear:[4]

(5) his, qui se dicunt
those.ABL who REFL.ACC say.3.PL
Judaeos esse, et non sunt
Jews.ACC be.INF and not be.3.PL
"those who say that they are Jews and are not"

(6) quia dicis quod dives sum . . . et
because say.2 DEM rich be.1 . . . and
nescis quia tu es miser
not.know.2 COMP you be.2 poor
"For you say, "I am rich, . . . " You do not realize that you are wretched"

(7) diabolus . . . sciens quod modicum
devil . . . knowing COMP short
tempus habet
time has.3
"the devil [has come down to you with great wrath], because he knows that his time is short"

Example 5 shows the Classical infinitive clause and 7 the Medieval *quod*-clause. Example 6 appears to be a transitional form, in which the first *quod* is not a complementizer, but a demonstrative introducing a direct quote ("you say *this*: I . . . "). This is evident from the following first-person verb, where an indirect quote ought to be in second person. The use of *quod* here echoes the Greek text and is an instance of the well-known influence of the Greek Bible on Christian Latin (Löfstedt, 1959, ch. 6).

8 Discussion

We find that TSGs are effective at identifying several historical changes in a modestly-sized corpus of Latin text. This extends the results of earlier papers which use TSGs to identify the writing of non-native English users. Here, the same features are applied to changes across time; we anticipate that similar results could be obtained in synchronic analysis of different dialects.

The approach does have significant limitations, however. Firstly, the dependence on treebank parses limits the set of texts to which the method can be applied. Parsing historical data may require specialized techniques (Pettersson et al., 2013) and fits within a larger set of cross-domain parsing problems which are notoriously difficult (McClosky et al., 2010). In particular, we suspect that the most difficult constructions will be precisely the ones which are novel in a particular era or region, since these may not appear in the training data. Parsers for Latin of any kind are rare, although working systems (McGillivray, 2013; Passarotti and Dell'Orletta, 2010) do exist.

Secondly, as seen above, the system has trouble unifying different examples of large constructions, such as clauses with and without modifiers. This prevents it from learning constructions larger than one or two context-free rules due to data sparsity. More expressive versions of TSG like Tree Adjoining Grammar (Joshi and Schabes, 1997) have been studied as solutions to this problem, including variants reducible to TSG (Swanson et al., 2013). It seems likely that such a more sophisticated grammatical representation could help to address this problem.

Although delexicalization of all content words was effective in controlling for the very different topics represented in our corpus, it also renders the system incapable of recognizing any lexically mediated changes. For instance, the system cannot represent changes in the argument structure

[4] Translations from the New Revised Standard Edition.

or subcategorization of a particular verb. Löfstedt (1959) lists changes such as datives with verbs of asking. Detecting this kind of change would require relexicalizing the trees, and therefore developing more sensitive statistical controls for topic. Due to the rarity of any individual word in a small corpus, however, a solution to this problem would be far less useful without methods for solving the previous ones as well. Only with a large automatically parsed corpus and a method for reducing fragmentations could enough examples of a lexically specific construction be gathered for any but the most common words.

Finally, the system cannot represent any changes involving semantic shifts. For instance, (Sidwell, 1995, p. 364) describes shifts in the tense system, including the use of pluperfect where perfect would be expected. Such changes cannot be detected from trees alone. Discovering them requires an ability to interpret the text and infer the implied time at which actions take place.

9 Conclusion

Despite these limitations, we believe TSGs offer a useful exploratory tool for discovering syntactic variation in corpora. Such a tool can allow historical linguists to learn about possible grammatical changes in dead languages for which they have no native intuition, broadening the kinds of questions they might investigate. This would parallel the recent use of computational systems to learn about lexical variation, allowing similar insights about the nature and history of syntactic change.

Acknowledgments

Ben Swanson, Brian Joseph, Alex Erdmann, Julia Papke and our reviewers, *gratias vobis agimus*.

References

J.H. Allen and J.B. Greenough. 1983. *Allen and Greenough's New Latin Grammar*. Caratzas Publishing Co., Inc., New Rochelle, New York.

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. A collaborative model of treebank development. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, pages 1–6.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Max Bane, Peter Graff, and Morgan Sonderegger. 2010. Longitudinal phonetic variation in a closed system. *Proc. CLS*, 46:43–58.

Rens Bod and Ronald Kaplan. 1998. A probabilistic corpus-driven model for lexical-functional analysis. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 145–151. Association for Computational Linguistics.

Trevor Cohn, Sharon Goldwater, and Phil Blunsom. 2009. Inducing compact but accurate tree-substitution grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 548–556. Association for Computational Linguistics.

Hal Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/, August.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *EACL*, pages 98–106.

Jacob Eisenstein. 2015. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Joshua Goodman. 1996. Efficient algorithms for parsing the DOP model. In *Proceedings of EMNLP*.

Jens T. Hanssen. 1945. Observations on Theodoricus Monachus and his history of the old Norwegian kings, from the end of the XII. sec. *Symbolae Osloenses*, 24.

Simon Hornblower, Antony Spawforth, and Esther Eidinow. 2012. *The Oxford Classical Dictionary*. Oxford University Press.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*.

Mark Johnson. 2002. The DOP estimation method is biased and inconsistent. *Computational Linguistics*, 28(1):71–76.

Aravind K Joshi and Yves Schabes. 1997. Tree-adjoining grammars. In *Handbook of formal languages*, pages 69–123. Springer.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 478–485, Barcelona, Spain, July.

Adam Ledgeway. 2012. *From Latin to Romance: Morphosyntactic typology and change*. Oxford University Press.

L.R. Lind. 1941. *Medieval Latin studies: Their nature and possibilities*. University of Kansas Publications.

Einar Löfstedt. 1959. *Late Latin*. Instituttet for Sammenlignende Kulturforsking.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California, June. Association for Computational Linguistics.

Barbara McGillivray. 2013. *Methods in Latin Computational Linguistics*. Brill.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 99–106. Association for Computational Linguistics.

Marco Passarotti and Felice Dell'Orletta. 2010. Improvements in parsing the Index Thomisticus treebank. revision, combination and a feature model for medieval Latin. In *Proceedings of LREC*.

Marco Carlo Passarotti. 2011. Language resources. the state of the art of Latin and the Index Thomisticus treebank project. In *Corpus anciens et Bases de donnes, ALIENTO. changes sapientiels en Mditerrane*, pages 301–320. ALIENTO.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, volume 18, pages 54–69.

Harm Pinkster. 1990. *Latin Syntax and Semantics*. Routledge.

Matt Post and Daniel Gildea. 2009. Bayesian learning of a tree substitution grammar. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 45–48. Association for Computational Linguistics.

Paolo Ramat. 1991. On Latin absolute constructions. *Linguistic Studies on Latin: Selected Papers from the 6th International Colloquium on Latin Linguistics*, pages 259–268.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Federico Sangati and Willem Zuidema. 2011. Accurate parsing with compact tree-substitution grammars: Double-DOP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 84–95. Association for Computational Linguistics.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86. Association for Computational Linguistics.

Graham Shipley, John Vandespoel, David Mattingly, and Lin Foxhall. 2008. *The Cambridge Dictionary of Classical Civilization*. Cambridge University Press.

Keith Sidwell. 1995. *Reading Medieval Latin*. University of Cambridge.

Ian Stewart. 2014. Now we stronger than ever: African-american syntax in Twitter. *EACL 2014*, page 31.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 193–197. Association for Computational Linguistics.

Ben Swanson and Eugene Charniak. 2013. Extracting the native language signal for second language acquisition. In *HLT-NAACL*, pages 85–94.

Ben Swanson and Eugene Charniak. 2014. Data driven language transfer hypotheses. In *EACL*, pages 169–173.

Ben Swanson, Elif Yamangil, Eugene Charniak, and Stuart M Shieber. 2013. A context free TAG variant. In *ACL (1)*, pages 302–310.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610. Association for Computational Linguistics.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 699–709. Association for Computational Linguistics.

F.A. Wright. 1933. *Letter to Eustochium: Select letters of St. Jerome*. Harvard University Press.